# Time-Limited and k-Limited Polling Systems: A Matrix Analytic Solution

Ahmad Al Hanbali, Roland de Haan, Richard J. Boucherie,
and Jan-Kees van Ommeren
University of Twente, Enschede, The Netherlands

October 5, 2009

**Abstract**

In this paper, we will develop a tool to analyze polling systems with the autonomous-server, the time-limited, and the k-limited service discipline. It is known that these disciplines do not satisfy the well-known branching property in polling system, therefore, hardly any exact result exists in the literature for them. Our strategy is to apply an iterative scheme that is based on relating in closed-form the joint queue-length at the beginning and the end of a server visit to a queue. These kernel relations are derived using the theory of absorbing Markov chains. Finally, we will show that our tool works also in the case of a tandem queueing network with a single server that can serve one queue at a time.

**Keywords:** Absorbing Markov chains; Matrix analytic solution; Polling system; Autonomous-server discipline; Time-limited discipline; $k$-limited discipline; Iterative scheme; Performance analysis;

## 1 Introduction

Polling systems have been extensively studied in the last years due to their vast area of applications in production and telecommunication systems [12, 16]. They have demonstrated to offer an adequate modeling framework to analyze systems in which a set of entities need certain service from a single resource. These entities are located at different positions in the system awaiting their turn to receive service.

In queueing theory, a polling system is equivalent to a set of queues with exogenous job arrivals all requiring an amount of service from a single server. The server serves each queue according to a specific service discipline and after serving a queue he will move to a next queue. A key role in the analysis of such polling systems is played by the so-called branching property [15]. This property states that each job present at a queue at the arrival instant of the

server will be replaced in an independent and indentically distributed manner by a random number of jobs during the course of the server's visit. Service disciplines satisfying the branching property yield a tractable analysis, while for disciplines not satisfying this property hardly any exact results are known.

The two most well-known disciplines that satisfy the bran- ching property are the exhaustive and gated discipline. Exhaustive means that the server continues servicing a queue until it becomes empty. At this instant the server moves to the next queue in his schedule. Gated means that the server only serves the jobs present in the queue at its arrival.

The drawback of the exhaustive and gated disciplines is that the server is controlled by the job arrivals. To reduce this control on the server, other type of service disciplines were introduced such as the time-limited and the $k$-limited discipline. According to the time-limited discipline, the server continues servicing a queue for a certain time period or until the queue becomes empty, whichever occurs first. Under the k-limited discipline, the server continues servicing a queue until $k$ jobs are served or the queue becomes empty, whichever occurs first. Another discipline, evaluated more recently in the literature and closely related to the time-limited discipline, is the so-called autonomous-server discipline [1, 4] which works as follows. The server continues servicing a queue for a certain period of time despite that, meanwhile, the queue may become empty. This discipline may also be seen as the non-exhaustive time-limited discipline. We should emphasize that these latter disciplines do not verify the branching property and thus hardly any closed-form results are known for the queue-length distribution under these disciplines.

To circumvent this difficulty, researchers resort to numerical methods using for instance iterative solution techniques or by using a power series algorithm. The power series algorithm [2, 3] aims at solving the global balance equations. To this end, the state probabilities are written as a power series and via a complex computation scheme the coefficients of these series, and thus the queue-length probabilities, are obtained. The iterative techniques [10, 11] exploit the relations between the joint queue-length distributions at specific instants, viz., the start of a server visit and the end of a server visit. The relation between the queue length at the start and end of a visit to a queue is established via recursively expressing the queue length at a job departure instant in terms of the queue length at the previous departure instant of a job. The complementary relation, between the queue length at the end of a visit to a queue and a start of visit to a next queue, can easily be established via the switch-over time. Starting with an initial distribution, the stationary queue-length distribution is then obtained by means of iteration. Although these methods offer a way to numerically solve intrinsically hard systems, their solution provides little fundamental insight and moreover the computation time and memory requirements to obtain this solution are exponential functions of the number of queues.

In this paper, we develop a tool to analyze the autonomous server, the time-limited, and the k-limited discipline. Our tool incorporates an iterative solution method which enhances the method introduced in [10]. More specifically, contrary to that approach, we will establish a direct and more insightful relation

between the joint number of jobs at the beginning and end of a visit period to a queue without conditioning on any intermediate events that occur during a visit. To this end, we use the theory of absorbing Markov chains (AMC) [9, 13]. We construct an AMC whose transient states represent the states of the polling system. The event of the server leaving a queue is modeled as an absorbing event. We will set the initial state of the AMC to the joint number of jobs at the beginning of a service period of a queue. Therefore, to find the joint number of jobs at the end of a service period, it is sufficient to keep track of the state from which the transition to the absorption state occurs. The probability of the latter event is eventually determined by first ordering the states in a careful way and consequently exploiting the structures that arise in the generator matrix of the AMC. Following this approach, we relate in closed-form the joint queue-length probability generating functions (p.g.f.) at the end of a visit period to a queue to the joint queue-length p.g.f. at the beginning of this visit period. The major part of this paper is devoted to deriving these kernel relations for the above-mentioned three disciplines: autonomous-server, time-limited, and k-limited. Once these relations are obtained, the joint queue-length distribution at server departure instants is readily obtained via a simple iterative scheme.

The paper is organized as follows. In Section 2 we give a careful description of the model and the assumptions. Section 3 analyses the autonomous-server discipline. In Section 4 we study the time-limited discipline. Section 5 evaluates the k-limited discipline. In Section 6 we describe the iterative scheme that is important to compute the joint queue-length distribution. Section 7 analyses briefly the tandem model case with the autonomous-server and the time-limited service discipline. Finally, in Section 8, we conclude the paper and give some research directions.

# 2 Model

We consider a single-server polling model consisting of $M$ first-in-first-out (FIFO) systems with unlimited queue, $Q_i$, $i = 1, \ldots, M$. Jobs arrive to $Q_i$ according to a Poisson process with arrival rate $\lambda_i$. We let $N_i(t)$ denote the number of jobs in $Q_i$, $i = 1, \ldots, M$, at time $t \geq 0$ and it is assumed that $N_i(0) = 0$, $i = 1, \ldots, M$. The service requirement $B_i$ at $Q_i$ has an exponential distribution $B_i(\cdot)$ and mean $b_i$. We assume that the service requirements are independent and identically distributed (iid) random variables (rvs). The server visits the queues in a cyclic fashion. After a visit to $Q_i$, the server incurs a switch-over time $C^i$ from $Q_i$ to $Q_{i+1}$. We assume that $C^i$ is independent of the service requirement and follows a general distribution $C^i(\cdot)$ with mean $c^i$, where at least one $c^i > 0$. The service discipline at each queue is either autonomous-server, time-limited, or $k$-limited. It is assumed that the queues of the polling system are stable.

In case the server is active at the end of a server visit, which may happen under the autonomous-server and time-limited disciplines, then the service will be preempted. At the beginning of the next visit of the server, the service time

will be re-sampled according to $B_i(\cdot)$. This discipline is commonly referred to as *preemptive-repeat-random*.

A word on notation. Given a random variable $X$, $X(t)$ will denote its distribution function. We use $\mathbf{I}$ to denote an identity matrix of appropriate size and use $\otimes$ as tensor product operator defined as follows. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices and $a(i,j)$ and $b(i,j)$ denote the (i,j)-entries of $\mathbf{A}$ and $\mathbf{B}$ respectively then $\mathbf{A} \otimes \mathbf{B}$ is a block matrix where the (i,j)-block is equal to b(i,j)$\mathbf{A}$. We use $e$ to denote a row vector of elements equal to one and $e_i$ to denote a row vector with the $i$-th element equal to one and the other elements equal to zero. Finally, $v^T$ will denote the transpose of vector $v$.

# 3 Autonomous-server discipline

In this section, we will relate the joint queue-length probabilities at the beginning and end of a server visit to a queue for the autonomous-server discipline. Under the autonomous-server discipline, the server remains at location $Q_i$ an exponentially distributed time with rate $\alpha_i$ before it migrates to the next queue in the cycle. It is stressed that even when $Q_i$ becomes empty, the server will remain at this queue.

Without loss of generality let us consider a server visit to $Q_1$. We assume that the p.g.f. of the steady-state queue-length at service's beginning instant at $Q_1$, denoted by $\beta_1^A(\mathbf{z})$, is known, where $\mathbf{z} := (z_1, \ldots, z_M)$ and $|z_i| \leq 1$ for $i = 1, \ldots, M$. The aim is to derive the p.g.f. of the steady-state queue-length at service visit's end at $Q_1$, denoted by $\gamma_1^A(\mathbf{z})$. In order to compute $\gamma_1^A(\mathbf{z})$, we first assume that $Q_1$ has a limited length of $L-1$ jobs including the job in service. This queue is denoted by $Q_1^L$. Later, we will let $L$ tend to infinity to get the desired results.

The probability that there are $(i_1, ..., i_M)$ jobs in $(Q_1, \ldots, Q_M)$ at the beginning of a server visit to $Q_1$ is denoted by $\mathbb{P}_L\big(\mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$. Similarly, the probability that there are $(j_1, ..., j_M)$ jobs in $(Q_1, \ldots, Q_M)$ at the end of a server visit to $Q_1$ is denoted by $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M) \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$. Under the assumption that the unlimited $Q_1$ is stable, $\lim_{L\to\infty} \mathbb{P}_L\big(\mathbf{N}_1^b = (i_1, \ldots, i_M)\big) = \mathbb{P}\big(\mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$ and $\beta_1^A(\mathbf{z}) = \mathbb{E}[\mathbf{z}^{\mathbf{N}_1^b}]$ .

Let $\mathbf{N}(t) := (N_1(t), \ldots, N_M(t))$ denote the $M$-dimensional, continuous-time Markov chain with discrete state-space $\xi_A = \{0, 1, \ldots, L-1\} \times \{0, 1, \ldots\}^{M-1} \cup \{a\}$, where $N_j(t)$ represents the number of jobs in $Q_j$ at time $t$. State $\{a\}$ is absorbing. We refer to this absorbing Markov chain by $\mathbf{AMC}_A$. The absorption of $\mathbf{AMC}_A$ occurs when the server leaves $Q_1$ which happens with rate $\alpha_1$. Moreover, the initial state of $\mathbf{AMC}_A$ at $t = 0$ is set to the system state at server's arrival to $Q_1$, i.e., $N_1^b = (i_1, \ldots, i_M)$. Therefore, the probability that the absorption of $\mathbf{AMC}_A$ occurs from one of the states $\{(j_1, \ldots, j_M)\}$ equals $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M) \mid N_1^b = (i_1, \ldots, i_M)\big)$. Let $\mathbf{n} = (n_1, \ldots, n_M) \in \xi_A - \{a\}$ and $e_l$ the $M$-dimensional row vector whose entries equal zero except the $l$-th entry that equals one. The non-zero transition rates of $\mathbf{AMC}_A$ can be written

as

$$\begin{array}{lcll} q(\mathbf{n}, \mathbf{n}+e_1) & = & \lambda_1, & 0 \leq n_1 \leq L-2, \\ q(\mathbf{n}, \mathbf{n}+e_l) & = & \lambda_l, & 2 \leq l \leq M, \\ q(\mathbf{n}, \mathbf{n}-e_1) & = & 1/b_1, & 1 \leq n_1 \leq L-1, \\ q\big(\mathbf{n}, \{a\}\big) & = & \alpha_1. & \end{array}$$

We derive now $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M) | \ \mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$. During a server visit to $Q_1$, the number of jobs at $Q_l$, $l = 2, \ldots, M$, may only increase. Therefore $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M\ ) \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$ is strictly positive for $j_l \geq i_l$, $l = 2, \ldots, M$, and zero otherwise. For sake of clarity, we will show first in detail the structure of $\mathbf{AMC}_A$ in the case of 3 queues, i.e. for $M = 3$, before considering the general case.

**Case M=3**. Let us consider the transient states of $\mathbf{AMC}_A$, i.e., $(n_1, n_2, n_3) \in \xi_A - \{a\}$, where $n_1 \in \{0, 1, \ldots, L-1\}$ and $n_2, n_3 \in \{0, 1, \ldots\}$. We recall that we consider a server visit to $Q_1$. The number of jobs at $Q_2$ and $Q_3$ may only increase during a server visit to $Q_1$, while the number of jobs at $Q_1$ may increase or decrease. To take advantage of this property, we will order the transient states of the $\mathbf{AMC}_A$ as follows: $(0,0,0), (1,0,0), (2,0,0), \ldots, (0,1,0), (1,1,0), (2,1,0), \ldots, (0,0,1), (1,0,1), (2,0,1), \ldots$, i.e., lexicographically orde- red first according to $n_3$, then $n_2$, and finally according to $n_1$. This ordering induces that the generator matrix of the transition rates between the transient states of $\mathbf{AMC}_A$ for $M = 3$, denoted by $\mathbf{Q}_3$, satisfies the following structure. That is, $\mathbf{Q}_3$ is an infinite upper-bidiagonal block matrix with diagonal blocks equal to $\mathbf{A}_3$ and upper-diagonal blocks equal $\lambda_3\mathbf{I}$, i.e.,

$$\mathbf{Q}_3 = \begin{pmatrix} \mathbf{A}_3 & \lambda_3\mathbf{I} & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_3 & \lambda_3\mathbf{I} & \mathbf{0} & \cdots \\ \vdots\ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1}$$

We note that $\mathbf{A}_3$ denotes the generator matrix of the transitions which do not induce any modification in the number of jobs at $Q_3$. Moreover, $\lambda_3\mathbf{I}$ denotes the transition rate matrix between the transient states $(n_1, n_2, n_3)$ and $(n_1, n_2, n_3 + 1)$, i.e., the transitions that represent an arrival to $Q_3$. The block matrix $\mathbf{A}_3$ is also an infinite upper-bidiagonal block matrix with diagonal blocks equal to $\mathbf{A}_2$, and upper-diagonal blocks equal $\lambda_2\mathbf{I}$, i.e.,

$$\mathbf{A}_3 = \begin{pmatrix} \mathbf{A}_2 & \lambda_2\mathbf{I} & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_2 & \lambda_2\mathbf{I} & \mathbf{0} & \cdots \\ \vdots\ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2}$$

where $\lambda_2\mathbf{I}$ denotes the transition rate matrix between the transient states $(n_1, n_2, n_3)$ and $(n_1, n_2 + 1, n_3)$ and $\mathbf{A}_2$ is the generator matrix of the transition between the transient states $(n_1, n_2, n_3)$ and $(n_1 \pm 1, n_2, n_3)$. Observe that $\mathbf{A}_2$ equals the sum of the generator matrix of an M/M/1/L-1 queue with arrival rate $\lambda_1$

and departure rate $1/b_1$ and of the matrix $-(\lambda_2+\lambda_3+\alpha_1)\mathbf{I}$. Now, we compute $\mathbb{P}_L\big(\mathbf{N}_1^e=(j_1,j_2,j_3)\mid \mathbf{N}_1^b=(i_1,i_2,i_3)\big)$ as function of the inverse of $\mathbf{Q}_3$, $\mathbf{A}_3$ and $\mathbf{A}_2$. First note that since $\mathbf{Q}_3$, $\mathbf{A}_3$ and $\mathbf{A}_2$ are all sub-generators with sum of their row elements strictly negative, these matrices are invertible. From the theory of absorbing Markov chains, given that $\mathbf{AMC}_A$ starts in state $(i_1,i_2,i_3)$, the probability that the transition to the absorption state $\{a\}$ occurs from state $(j_1,j_2,j_3)$ reads (see, e.g., [8])

$$\mathbb{P}_L\big(\mathbf{N}_1^e=(j_1,j_2,j_3)\mid \mathbf{N}_1^b\big)=-\alpha_1 c_3(\mathbf{Q}_3)^{-1}d_3, \tag{3}$$

where $c_3$ is the probability distribution vector of $\mathbf{AMC}_A$'s initial state that can be given by

$$c_3:=e_{i_1}\otimes e_{i_2}\otimes e_{i_3},$$

and $\alpha_1 d_3$ is the transition rate vector to $\{a\}$ given that $(j_1,j_2,j_3)$ is the last state visited before absorption where $d_3$ can be given by

$$d_3:=(e_{j_1}\otimes e_{j_2}\otimes e_{j_3})^T.$$

$\mathbf{Q}_3$ is an upper-bidiagonal block matrix. Hence, it is easy to show that $(\mathbf{Q}_3)^{-1}$ is an upper-triangular block matrix with (i,j)-block equal to $(-(\mathbf{A}_3)^{-1}\lambda_3\mathbf{I})^{j-i}(\mathbf{A}_3)^{-1}$, thus we find that

$$c_3(\mathbf{Q}_3)^{-1}d_3 \quad = \quad c_2(-\lambda_3(\mathbf{A_3})^{-1})^{j_3-i_3}(\mathbf{A}_3)^{-1}d_2, \tag{4}$$

where $c_2=e_{i_1}\otimes e_{i_2}$ and $d_2=(e_{j_1}\otimes e_{j_2})^T$. Plugging (4) into (3) gives that

$$\mathbb{P}_L\big(\mathbf{N}_1^e=(j_1,j_2,j_3)\mid \mathbf{N}_1^b\big)=-\alpha_1 c_2(-\lambda_3(\mathbf{A_3})^{-1})^{j_3-i_3}(\mathbf{A}_3)^{-1}d_2. \tag{5}$$

**General case**. By analogy with the case of $M=3$, we order the transient states of $\mathbf{AMC}_A$ first according to $n_M$, then $n_{M-1}$, ..., and finally according to $n_1$. During a server visit to $Q_1$, the number of jobs at $Q_j$, $j=2,\ldots,M$, may only increase. Therefore, similarly to the case of $M=3$, the $\mathbf{AMC}_A$ the generator matrix of the transition rates between the transient states of $\mathbf{AMC}_A$ for the general case, denoted by $\mathbf{Q}_M$, is an upper-bidiagonal block matrix with diagonal blocks equal to $\mathbf{A}_M$, and upper-diagonal blocks equal to $\lambda_M\mathbf{I}$. Moreover, $\mathbf{A}_M$ in turn is an upper-bidiagonal block matrix with diagonal blocks equal to $\mathbf{A}_{M-1}$, and upper-diagonal blocks equal to $\lambda_{M-1}\mathbf{I}$. We emphasize that $\mathbf{A}_j$, $j=M,\ldots,3$, all verify the previous property. Finally, the matrix $\mathbf{A}_2$ equals the sum of the generator matrix of an M/M/1/L-1 queue with arrival rate $\lambda_1$ and departure rate $1/b_1$ and of the matrix $-(\lambda_2+\ldots+\lambda_M+\alpha_1)\mathbf{I}$.

By analogy with the $M=3$ case, we find that the probability of $\mathbf{N}_i^e=(j_1,\ldots,j_M)$, given that $\mathbf{N}_1^b=(i_1,\ldots,i_M)$, reads

$$\mathbb{P}_L\big(\mathbf{N}_1^e=(j_1,\ldots,j_M)\mid \mathbf{N}_1^b=(i_1,\ldots,i_M)\big)= \\ -\alpha_1 c_{M-1}\big(-\lambda_M(\mathbf{A_M})^{-1}\big)^{j_M-i_M}(\mathbf{A_M})^{-1}d_{M-1}. \tag{6}$$

$$\begin{aligned} c_{M-1} &:= e_{i_1} \otimes \ldots \otimes e_{i_{M-1}}, \\ d_{M-1} &:= (e_{j_1} \otimes \ldots \otimes e_{j_{M-1}})^T. \end{aligned}$$

We derive now the conditional p.g.f. of $\mathbf{N}_1^e$. Note that $\big(-\lambda_M(\mathbf{A}_M)^{-1}\big)$ is a sub-stochastic matrix with the sum of its row elements strictly smaller than one, which gives that $\lim_{n\to\infty}(-\lambda_M(\mathbf{A}_M)^{-1})^n = \mathbf{0}$. Combining the latter result with (6) we find that

$$\mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] = -\alpha_1 z_M^{i_M} c_{M-1}\big(\mathbf{A}_M + z_M\lambda_M\mathbf{I}\big)^{-1} d_{M-1}(\mathbf{z}), \quad (7)$$

where

$$d_{M-1}(\mathbf{z}) := \sum_{j_1=0}^{L-1} \sum_{j_2 \geq i_2} \ldots \sum_{j_{M-1} \geq i_{M-1}} (z_1^{j_1} e_{j_1} \otimes \ldots \otimes z_{M-1}^{j_{M-1}} e_{j_{M-1}})^T, \quad (8)$$

and $|z_i| \leq 1$, $i = 1, \ldots, M$. It remains to find $(\mathbf{A}_M + z_M\lambda_M\mathbf{I})^{-1}$. Since $\mathbf{A}_M$ is an upper-bidiagonal block matrix, the (i,j)-block of $(\mathbf{A}_M + z_M\lambda_M\mathbf{I})^{-1}$ is given by $(-\lambda_{M-1})^{j-i}\times$ $(\mathbf{A}_{M-1} + z_M\lambda_M\mathbf{I})^{-j+i-1}$. Plugging the latter result into (7) gives that

$$\begin{aligned} E_L\Big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] &= -\alpha_1 z_M^{i_M} z_{M-1}^{i_{M-1}} c_{M-2} \times \\ &\big(\mathbf{A}_{M-1} + (z_M\lambda_M + z_{M-1}\lambda_{M-1})\mathbf{I}\big)^{-1} d_{M-2}(\mathbf{z}), \quad (9) \end{aligned}$$

where

$$\begin{aligned} c_{M-2} &:= e_{i_1} \otimes \ldots \otimes e_{i_{M-2}}, \\ d_{M-2}(\mathbf{z}) &:= \sum_{j_1=0}^{L-1} \sum_{j_2 \geq i_2} \ldots \sum_{j_{M-2} \geq i_{M-2}} (z_1^{j_1} e_{j_1} \otimes \ldots \otimes z_{M-2}^{j_{M-2}} e_{j_{M-2}})^T. \end{aligned}$$

By an induction argument along with the properties that $\mathbf{A}_j$, $j = 3, \ldots, M-1$, is an upper-bidiagonal block matrix, it can be shown that

$$\begin{aligned} \mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] &= -\alpha_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1} \times \\ &\Big(\mathbf{A}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}\Big)^{-1} d_1(z_1), \quad (10) \end{aligned}$$

where

$$d_1(z_1) := \sum_{j_1=0}^{L-1} z_1^{j_1}(e_{j_1})^T = (1, z_1, \ldots, z_1^{L-1})^T.$$

Removing the condition on $\mathbf{N}_1^b$, it is readily seen that

$$\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e}\big] = -\alpha_1 f\Big(\mathbf{A}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}\Big)^{-1} d_1(z_1), \quad (11)$$

where $f$ is the $L$-dimensional row vector with $i$-th element equal to $\mathbb{E}\big[\mathbf{1}_{\{N_1^b=i\}} \cdot z_2^{N_2^b} \ldots z_M^{N_M^b}\big]$, for $i = 0, \ldots, L-1$. It remains to find the inverse of $\mathbf{A}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}$ and to let $L \to \infty$.

Let $u^T = (1, 0, \ldots, 0)$ and let $v^T = (0, \ldots, 0, 1)$. We recall that $\mathbf{A}_2$ equals the sum of the generator matrix of an M/M/1/L-1 queue with arrival rate $\lambda_1$ and departure rate $1/b_1$ and of the matrix $-(\lambda_2 + \ldots + \lambda_M + \alpha_1)\mathbf{I}$. Let $\mathbf{Q_A}(\mathbf{z}) := \mathbf{A}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}$. Now, observe that $\mathbf{Q_A}(\mathbf{z}) = \mathbf{T_A}(\mathbf{z}) + 1/b_1 uu^T + \lambda_1 vv^T$, where $\mathbf{T_A}(\mathbf{z})$ is a L-by-L tridiagonal Toeplitz matrix with diagonal entries equal $\big(-\lambda_1 - 1/b_1 - \alpha_1 - \sum_{m=2}^M \lambda_m(1-z_m)\big)$, upper-diagonal entries equal $\lambda_1$, and lower-diagonal entries $1/b_1$. Let $t_{ij}^*$ denote the $(i,j)$-entry of $\mathbf{T_A^{-1}}(\mathbf{z})$. By applying the Sherman-Morrison formula [14, p. 76] we find that the $(i,j)$-entry of $\mathbf{Q_A^{-1}}(\mathbf{z})$ gives for $i, j = 1, \ldots, L$,

$$q_{ij}^* = m_{ij} - \lambda_1 \frac{m_{iL} m_{Lj}}{1 + \lambda_1 m_{LL}}, \quad \text{where } m_{ij} = t_{ij}^* - \frac{t_{i1}^* t_{1j}^*}{b_1 + t_{11}^*}. \tag{12}$$

The inverse of a tridiagonal Toeplitz matrix is known in closed-form (see [5, Sec. 3.1])

$$t_{ij}^* = \begin{cases} -\frac{(r_{11}^i - r_{21}^i)(r_{11}^{L+1-j} - r_{21}^{L+1-j})}{\lambda_1(r_{11}-r_{21})(r_{11}^{L+1} - r_{21}^{L+1})} & , \ i \le j \le L \\ \frac{(r_{11}^{-j} - r_{21}^{-j})(r_{11}^{L+1-i} - r_{21}^{L+1-i})}{\lambda_1(r_{11}-r_{21})(r_{11}^{L+1} - r_{21}^{L+1})} & , \ j \le i \le L \end{cases} \tag{13}$$

where $r_{11}$ and $r_{21}$ are the distinct roots of

$$P_1(r) := \lambda_1 r^2 - s_1 r + 1/b_1, \tag{14}$$

where $s_1 := \lambda_1 + 1/b_1 + \alpha_1 + \sum_{m=2}^M \lambda_m(1 - z_m)$. We take $|r_{11}| < |r_{21}|$. Note that $|\lambda_1 r^2 + 1/b_1| < |-s_1 r|$ for every $|r| = 1$, thus Rouché's theorem gives that $P_1(r)$ has exactly one root inside the disk of radius one for all $|z_i| \le 1$ (see, e.g., [7]). For this reason, we have that $|r_{11}| < 1 < |r_{21}|$.

Inserting the values of $t_{ij}^*$ into (11) yields that

$$\begin{aligned} \mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e}\big] &= -\alpha_1 \sum_{i=0}^{L-1} f(i) \sum_{j=1}^{L} z_1^{j-1} \Big[ t_{ij}^* - \frac{1/b_1 t_{i1}^* t_{1j}^*}{1 + 1/b_1 t_{11}^*} \\ &\quad - \frac{\lambda_1 m_{iL}}{1 + \lambda_1 m_{LL}} \Big( t_{Lj}^* - \frac{1/b_1 t_{L1}^* t_{1j}^*}{1 + 1/b_1 t_{11}^*} \Big) \Big]. \end{aligned} \tag{15}$$

Thus, it remains to let $L \to \infty$ in (15) in order to find $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big]$. It is readily seen that

$$\begin{aligned} \lim_{L\to\infty} t_{LL-j}^* &= -\frac{1}{\lambda_1 r_{21}} r_{11}^j, \\ \lim_{L\to\infty} m_{L-iL} &= \lim_{L\to\infty} t_{L-iL}^* = -\frac{1}{\lambda_1} r_{21}^{-(i+1)}, \\ \lim_{L\to\infty} t_{1j}^* &= -\frac{1}{\lambda_1} r_{21}^{-j}, \\ \lim_{L\to\infty} t_{i1}^* &= \frac{-1}{\lambda_1 r_{11} r_{21}} r_{11}^i. \end{aligned}$$

Some technical calculus shows that the following limit is equal to zero

$$\lim_{L\to\infty} \alpha_1 \sum_{i=0}^{L-1} f(i) \sum_{j=1}^{L} z_1^{j-1} \frac{\lambda_1 m_{iL}}{1+\lambda_1 m_{LL}} \Big( t_{Lj} - \frac{1/b_1 t_{L1} t_{1j}}{1+1/b_1 t_{11}} \Big).$$

Finally, plugging the previous limits in (15) it can be shown that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big] = \gamma_1^A(\mathbf{z}) = \frac{\alpha_1(1-z_1)}{P_1(z_1)} \Big( \frac{r_{11}\beta_1^A(\mathbf{z}_1^*)}{1-r_{11}} - \frac{z_1\beta_1^A(\mathbf{z})}{1-z_1} \Big), \tag{16}$$

where $\mathbf{z}_1^* := (r_{11}, z_2, \ldots, z_M)$. Eq. (16) relates in closed-form $\gamma_1^A(\mathbf{z})$, p.g.f. of the joint queue-length at the beginning of a server visit to $Q_1$, to $\beta_1^A(\mathbf{z})$, p.g.f. of the joint queue-length at the end of a server visit to $Q_1$. From (16), we deduce that for a server visit to $Q_i$, $i = 1, \ldots, M$,

$$\gamma_i^A(\mathbf{z}) = \frac{\alpha_i(1-z_i)}{P_i(z_i)} \Big( \frac{r_{1i}\beta_i^A(\mathbf{z}_i^*)}{1-r_{1i}} - \frac{z_i\beta_i^A(\mathbf{z})}{1-z_i} \Big), \tag{17}$$

where

$$P_i(z_i) := \lambda_i z_i^2 - s_i z_i + 1/b_i, \tag{18}$$

$$s_i := \lambda_i + 1/b_i + \alpha_i + \sum_{m=1, m\neq i}^{M} \lambda_m(1-z_m), \tag{19}$$

$$r_{1i} := \frac{s_i - \sqrt{(s_i)^2 - 4\lambda_i/b_i}}{2\lambda_i}, \tag{20}$$

$\mathbf{z}_i^* := (z_1, \ldots, z_{i-1}, r_{1i}, z_{i+1}, \ldots, z_M)$, and $|r_{1i}| < 1$.

Finally, introducing the switch-over times from $Q_{i-1}$ to $Q_i$, thus by using that $\beta_i^A(\mathbf{z}) = \gamma_{i-1}^A(\mathbf{z}) C^{i-1}(\mathbf{z})$, where $C^{i-1}(\mathbf{z})$ is the p.g.f. of the number of Poisson arrivals during $C^{i-1}$, we obtain

$$\begin{aligned} \gamma_i^A(\mathbf{z}) &= \frac{\alpha_i(1-z_i)r_{1i}}{P_i(z_i)(1-r_{1i})} \gamma_{i-1}^A(\mathbf{z}_i^*) C^{i-1}(\mathbf{z}_i^*) \\ &\quad - \frac{\alpha_i z_i}{P_i(z_i)} \gamma_{i-1}^A(\mathbf{z}) C^{i-1}(\mathbf{z}). \end{aligned} \tag{21}$$

# 4 Time-Limited discipline

In this section, we will relate the joint queue-length probabilities at the beginning and end of a server visit to a queue for the time-limited discipline. Under this discipline, the server departs from $Q_i$ when it becomes empty or when a timer of exponential distribution duration with rate $\alpha_i$ has expired, whichever occurs first. Moreover, if the server arrives to an empty queue, he leaves the queue immediately and jumps to the next queue in the schedule. For this reason, we should differentiate here between the two events where the server join an empty and non-empty queue.

We will follow the same approach as in Section 3. Thus, we first assume that $Q_1$ has a limited queue of $L-1$ jobs, second there are $\mathbf{N}_1^b := (i_1, ..., i_M)$ jobs in $(Q_1, \ldots, Q_M)$, with $i_1 \geq 1$, at the beginning time of a server visit to $Q_1$ and third there are $\mathbf{N}_1^e := (j_1, ..., j_M)$ jobs in $(Q_1, \ldots, Q_M)$ at the end time of a server visit to $Q_1$. Note that if $Q_1$ is empty at the beginning of a server visit, i.e., $i_1 = 0$, $\mathbb{P}\big(\mathbf{N}_1^e = \mathbf{N}_1^b\big) = 1$. We will exclude the latter obvious case from the analysis in the following, however, we will include it when we will uncondition on $\mathbf{N}_1^b$.

Let $\mathbf{N}(t) := (N_1(t), \ldots, N_M(t))$ denote the $M$-dimensional, continuous-time Markov chain with discrete state-space $\xi_T = \{1, \ldots, L-1\} \times \{0, 1, \ldots\}^{M-1} \cup \{a\}$, where $N_j(t)$ represents the number of jobs in $Q_j$ at time $t$ and at which $Q_1$ is being served. State $\{a\}$ is absorbing. We refer to this absorbing Markov chain by $\mathbf{AMC}_T$. The absorption of $\mathbf{AMC}_T$ occurs when the server leaves $Q_1$ which happens with rate $\alpha_1$ from all transient states. The transient states of the form $(1, n_2, \ldots, n_M)$ have an additional transition rate to $\{a\}$ that is equal to $1/b_1$, which represents the departure of the last job at $Q_1$.

We set $\mathbf{N}(0) = \mathbf{N}_1^b$. Therefore, the probability that the absorption of $\mathbf{AMC}_T$ occurs from one of the states $\{(j_1, \ldots, j_M)\}$ equals $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M)\big)$, if the absorption is due to the timer expiration with rate $\alpha_1$. However, if the absorption is due to $Q_1$ becoming empty, $\mathbb{P}_L\big(\mathbf{N}_1^e = (0, j_2 \ldots, j_M)\big)$ equals the probability that the absorption with rate $1/b_1$ occurs from one of the states $\{(1, j_2, \ldots, j_M)\}$. The non-zero transition rates of $\mathbf{AMC}_T$ can be written for all $\mathbf{n} \in \xi_T - \{a\}$,

$$\begin{array}{lcl} q(\mathbf{n}, \mathbf{n} + e_1) & = & \lambda_1, \quad n_1 = 1, \ldots, L-2, \\ q(\mathbf{n}, \mathbf{n} + e_l) & = & \lambda_l, \quad l = 2, \ldots, M, \\ q(\mathbf{n}, \mathbf{n} - e_1) & = & 1/b_1, \quad 2 \leq n_1 \leq L-1, \\ q\big(\mathbf{n}, \{a\}\big) & = & \alpha_1, \quad 2 \leq n_1 \leq L-1, \\ q\big(\mathbf{n}, \{a\}\big) & = & \alpha_1 + 1/b_1, \quad n_1 = 1. \end{array}$$

We derive now $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M) \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$. We order the transient states lexicographically first according to $n_M$, then to $n_{M-1}$, ..., and finally to $n_1$. Similarly to the time-limited discipline, during a server visit to $Q_1$, the number of jobs at $Q_j$, $j = 2, \ldots, M$, may only increase. It then follows that the transient generator of $\mathbf{AMC}_T$ has the same structure as the transient generator of $\mathbf{AMC}_A$, i.e. it is an upper-bidiagonal Toeplitz matrix of upper-bidiagonal Toeplitz diagonal blocks. Therefore, by the same arguments as for the time-limited discipline, we find that the joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to timer expiration, denoted by {timer}, given $\mathbf{N}_1^b$, reads

$$\mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e} \cdot \mathbf{1}_{\{\text{timer}\}} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] = \\ - \alpha_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1} \Big(\mathbf{B}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}\Big)^{-1} g_1(z_1),$$

where $\mathbf{B}_2$ is the sum of the generator matrix of an M/M/1/L-1 queue with arrival rate $\lambda_1$ and service rate $1/b_1$ restricted to the states with the number of

jobs strictly positive, and of the matrix $-(\lambda_2 + \ldots + \lambda_M + \alpha_1)\mathbf{I}$, and where

$$g_1(z_1) \quad := \quad (z_1, \ldots, z_1^{L-1})^T.$$

Let,

$$\mathbf{Q_T}(\mathbf{z}) := \mathbf{B}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}. \tag{22}$$

The joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to empty $Q_1$, denoted by $\{Q_1 \text{ empty}\}$, given $\mathbf{N}_1^b$, reads

$$\mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e} \cdot \mathbf{1}_{\{Q_1 \text{ empty}\}} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] = -1/b_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1} \big(\mathbf{Q_T}(\mathbf{z})\big)^{-1} e_1,$$

Summing the latter two p.g.f. gives the p.g.f. of $\mathbf{N}_1^e$ given $\mathbf{N}_1^b$, which reads

$$\begin{aligned}\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\big] \quad &= \quad -\alpha_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1} \times \\ &\big(\mathbf{Q_T}(\mathbf{z})\big)^{-1}\Big(g_1(z_1) + \frac{1}{b_1\alpha_1} \cdot e_1\Big), \end{aligned} \tag{23}$$

In the final part of this section, we find the inverse of $\mathbf{Q_T}(\mathbf{z})$ and let $L \to \infty$.

We note that $\mathbf{Q_T}(\mathbf{z}) = \mathbf{T}(\mathbf{z}) + \lambda_1 v v^T$, $v = (0, \ldots, 0, 1)^T$, where $\mathbf{T_T}(\mathbf{z})$ is a (L-1)-by-(L-1) tridiagonal Toeplitz matrix with diagonal entries equal to $\big(-\lambda_1 - 1/b_1 - \alpha_1 - \sum_{m=2}^{M} \lambda_m(1 - z_m)\big)$, upper-diagonal entries are equal to $\lambda_1$, and lower-diagonal entries $1/b_1$. We emphasize that the only difference between $\mathbf{T_A}(\mathbf{z})$ of the autonomous-server discipline and $\mathbf{T_T}(\mathbf{z})$ is that $\mathbf{T_A}(\mathbf{z})$ is an L-by-L matrix. Therefore, following the same approach as in Section 3, we find that the $(i,j)$-entry of $\mathbf{Q_T}(\mathbf{z})^{-\mathbf{1}}$, $i, j = 1, \ldots, L-1$, gives

$$q(i,j)^* = t(i,j)_T^* - \lambda_1 \frac{t(i, L-1)_T^* t(L-1, j)_T^*}{1 + \lambda_1 t(L-1, L-1)_T^*}, \tag{24}$$

where $t(i,j)_T^*$ is the (i,j)-entry of $\mathbf{T_T}(\mathbf{z})^{-\mathbf{1}}$ that reads

$$t(i,j)_T^* = \begin{cases} -\frac{(r_{11}^i - r_{21}^i)(r_{11}^{L-j} - r_{21}^{L-j})}{\lambda_1(r_{11} - r_{21})(r_{11}^L - r_{21}^L)} & , \ i \le j \le L-1 \\ \frac{(r_{11}^{-j} - r_{21}^{-j})(r_{11}^L r_{21}^i - r_{21}^L r_{11}^i)}{\lambda_1(r_{11} - r_{21})(r_{11}^L - r_{21}^L)} & , \ j \le i \le L-1 \end{cases} \tag{25}$$

where $r_{11}$ and $r_{21}$ are the distinct roots of $P_1(r) := \lambda_1 r^2 - s_1 r + 1/b_1$. Inserting the values of $q(i,j)_T^*$ into (23) yields that

$$\begin{aligned}\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\big] &= -\alpha_1 z_2^{i_2} \ldots z_M^{i_M} \times \\ &\left[\frac{1}{b_1\alpha_1} q(i_1, 1)^* + \sum_{j=1}^{L-1} z_1^j \left(t(i_1, j)_T^* - \lambda_1 \frac{t(i_1, L-1)_T^* \cdot t(L-1, j)_T^*}{1 + \lambda_1 t(L-1, L-1)_T^*}\right)\right]. \end{aligned} \tag{26}$$

Some technical calculus shows that the following limit is equal to zero

$$\lim_{L \to \infty} \frac{t(i_1, L-1)_T^*}{1 + \lambda_1 t(L-1, L-1)_T^*} \sum_{j=1}^{L-1} z_1^j \cdot t(L-1, j)_T^*.$$

Plugging the latter limit, $q(i_1,1)^*$, and $t(i_1,j)^*_T$ in (26), we find that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\ |\mathbf{N}_1^b=(i_1,\ldots,i_M)\ \big]=z_2^{i_2}\ldots z_M^{i_M}\left(r_{11}^{i_1}-\frac{\alpha_1 z_1}{P_1(z_1)}(z_1^{i_1}-r_{11}^{i_1})\right) \tag{27}$$

Removing the condition of $\mathbf{N}_1^b=(i_1,\ldots,i_M)$ for $i_1=0,\ldots,L-1$,

$$\gamma_1^T(\mathbf{z}) \quad = \quad \Big(1+\frac{\alpha_1 z_1}{P_1(z_1)}\Big)\beta_1^T(\mathbf{z}_1^*)-\frac{\alpha_1 z_1}{P_1(z_1)}\beta_1^T(\mathbf{z}), \tag{28}$$

where $\mathbf{z}_1^*:=(r_{11},z_2,\ldots,z_M)$. From (28), we deduce that for a server visit to $Q_i$, $i=1,\ldots,M$,

$$\gamma_i^T(\mathbf{z})=\Big(1+\frac{\alpha_i z_i}{P_i(z_i)}\Big)\beta_i^T(\mathbf{z}^*)-\frac{\alpha_i z_i}{P_i(z_i)}\beta_i^T(\mathbf{z}), \tag{29}$$

where $\mathbf{z}_i^*=(z_1,\ldots,z_{i-1},r_{1i},z_{i+1},\ldots,z_M)$, $|r_{1i}|<1$, and where $P_i(z_i)$, $s_i$, and $r_{1i}$ are in (18), (19), and (20) respectively.

Finally, introducing the switch-over times from $Q_{i-1}$ to $Q_i$, we obtain

$$\gamma_i^T(\mathbf{z}) \quad = \quad \Big(1+\frac{\alpha_i z_i}{P_i(z_i)}\Big)\gamma_{i-1}^T(\mathbf{z}_i^*)C^{i-1}(\mathbf{z}_i^*)-\frac{\alpha_i z_i}{P_i(z_i)}\gamma_{i-1}^T(\mathbf{z})C^{i-1}(\mathbf{z}). \tag{30}$$

# 5 k-Limited Discipline

In this section, we analyze the $k$-limited discipline. According to this discipline the server continues working at a queue until either a predefined number of $k$ jobs is served or the queue becomes empty, whichever occurs first. Similarly to the previous disciplines, the objective is to relate the joint queue-length probabilities at the beginning and end of a server visit to $Q_1$, referred to as $\beta_1^k(\mathbf{z})$ and $\gamma_1^k(\mathbf{z})$.

By analogy with the time-limited discipline, we will first assume that $Q_1$ has a limited queue of $L-1$ jobs, second there are $\mathbf{N}_1^b:=(i_1,...,i_M)$ jobs in $(Q_1,\ldots,Q_M)$, with $i_1\geq 1$, at the beginning time of a server visit to $Q_1$, and third there are $\mathbf{N}_1^b:=(j_1,...,j_M)$ jobs in $(Q_1,\ldots,Q_M)$ at the end time of a server visit to $Q_1$. Note that if $Q_1$ is empty at the beginning of a server visit, $i_1=0$, the server will leave immediately, i.e., $\mathbb{P}\big(\mathbf{N}_1^e=\mathbf{N}_1^b\big)=1$. For this reason, we will exclude the latter obvious case from the analysis in the following, however, we will include it when we will uncondition on $\mathbf{N}_1^b$.

Let $\mathbf{N}(t):=(N_1(t),\ldots,N_M(t),D(t))$ denote the $M+1$-dimensional, continuous-time Markov chain with discrete state-space $\xi_k=\{1,\ldots,L-1\}\times\{0,1,\ldots\}^{M-1}\times\{0,1,\ldots\}\cup\{a\}$, where $N_j(t)$ represents the number of jobs in $Q_j$ at time $t$ during a server visit to $Q_1$, and $D(t)$ is the total number of departures from $Q_1$ until $t$. State $\{a\}$ is absorbing. This absorbing Markov chain is denoted by $\mathbf{AMC}_k$. The absorption of $\mathbf{AMC}_k$ occurs when the server leaves $Q_1$ which happens with rate $1/b_1$ from all transient states with $D(t)=k-1$ or $N_1(t)=1$.

We set $\mathbf{N}(0)=(\mathbf{N}_1^b,0)$. The probability that the transition to the absorption state occurs from one of the states $\{(j_1,\ldots,j_M)\}$, $j_1\geq 2$, equals $\mathbb{P}_L\big(\mathbf{N}_1^e=(j_1-1,\ldots,j_M)\mid\mathbf{N}_1^b\big)$ and the absorption is eventually due to $k$ departures from

$Q_1$ with rate $1/b_1$. If the absorption is due to $Q_1$ becoming empty, $\mathbb{P}_L\big(\mathbf{N}_1^e = (0, j_2 \ldots, j_M) \mid \mathbf{N}_1^b\big)$ equals the probability that the transition to absorption is with rate $1/b_1$ and it occurs from state $\{(1, j_2, \ldots, j_M)\}$. Note that it is possible that the k-th departure at $Q_1$ leaves behind an empty queue. In our analysis we will consider this event as a transition to absorption that is due to $Q_1$ becoming empty. The non-zero transition rates of $\mathbf{AMC}_k$ can be written for all $\mathbf{n} = (n_1, \ldots, n_M, j) \in \xi_k - \{a\}$,

$$\begin{array}{lcll} q(\mathbf{n}, \mathbf{n} + e_1) & = & \lambda_1, & n_1 = 1, \ldots, L-2, \\ q(\mathbf{n}, \mathbf{n} + e_l) & = & \lambda_l, & l = 2, \ldots, M, \\ q(\mathbf{n}, \mathbf{n} - e_1 + e_{M+1}) & = & 1/b_1, & n_1 = 2, \ldots, L-1, \\ & & & j = 0, \ldots, k-2, \\ q(\mathbf{n}, \{a\}) & = & 1/b_1, & n_1 = 1 \text{ or } j = k-1. \end{array}$$

We derive now $\mathbb{P}_L\big(\mathbf{N}_1^e = (j_1, \ldots, j_M) \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\big)$. We order the transient states of $\mathbf{AMC}_k$ lexicographically according to $n_M$, $n_{M-1}$, ..., $n_2$, then to $j$, and finally according to $n_1$. During a server visit to $Q_1$, the number of jobs at $Q_j$, $j = 2, \ldots, M$, may only increase. Therefore, similarly to the automous-server and time-limited discipline, we deduce that the joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due $k$ to departures, denoted by {k dep.}, given $\mathbf{N}_1^b$, reads

$$\mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{\text{k dep.}\}} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] = -1/b_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1} \otimes e_1 \times \Big(\mathbf{C}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}\Big)^{-1} h(z_1), \quad (31)$$

where $e_1$ is a k-dimensional row vector of zero entries except the first that is one, $(\mathbf{C}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I})$ is a k-by-k upper-bidiagonal block matrix of upper diagonal blocks equal to $\mathbf{U}$, where $\mathbf{U}$ is an (L-1)-by-(L-1) lower-diagonal matrix whose entries equal to $1/b_1$, and of diagonal blocks equal to $\mathbf{D}$, where $\mathbf{D}$ is the sum of the generator matrix of a M/M/1/L-1 queue with arrival rate $\lambda_1$ and service rate 0 restricted to strictly positive states, and of the matrix $-(\lambda_2(1 - z_2) + \ldots + \lambda_M(1 - z_M) + 1/b_1)\mathbf{I}$, and

$$\begin{array}{rcll} h(z_1) & := & q(z_1) \otimes e_k, & (32) \\ q(z_1) & := & (0, z_1, \ldots, z_1^{L-2})^T, & (33) \end{array}$$

and where $e_k$ is a k-dimensional column vector of zero entries except the k-th that is one. Plugging the inverse of $\Big(\mathbf{C}_2 + (z_2\lambda_2 + \ldots + z_M\lambda_M)\mathbf{I}\Big)$ into (31) gives that

$$\mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e} \cdot \mathbf{1}_{\{\text{k dep.}\}} \mid \mathbf{N}_1^b = (i_1, \ldots, i_M)\Big] = -1/b_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1} \times \big(-\mathbf{D}^{-1}\mathbf{U}\big)^{k-1}\mathbf{D}^{-1} q(z_1), \quad (34)$$

The joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to empty $Q_1$, denoted by $\{Q_1 \text{ emp}\}$, given $\mathbf{N}_1^b$, reads

$$\begin{aligned}
&\mathbb{E}_L\Big[\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{Q_1 \text{ emp.}\}} \mid \mathbf{N}_1^b = (i_1,\ldots,i_M)\Big] \\
&= -1/b_1 z_2^{i_2}\ldots z_M^{i_M} e_{i_1}\otimes e_1\Big(\mathbf{C}_2 + (z_2\lambda_2+\ldots+z_M\lambda_M)\mathbf{I}\Big)^{-1} e_1\otimes e \\
&= -1/b_1 z_2^{i_2}\ldots z_M^{i_M} e_{i_1}\big(\mathbf{I}-\big(-\mathbf{D}^{-1}\mathbf{U}\big)^k\big)\big(\mathbf{D}+\mathbf{U}\big)^{-1} e_1. \qquad (35)
\end{aligned}$$

Summing the latter two p.g.f. gives $\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big]$. It remains to find first $e_{i_1}\big(-\mathbf{D}^{-1}\mathbf{U}\big)^k$, second $\big(\mathbf{D}^{-1}\big)q(z_1)$ and $\big(\mathbf{D}+\mathbf{U}\big)^{-1}e_1$, so that finally we will take the limit for $L\to\infty$ of $\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big]$ .

## 5.1 $e_{i_1}\big(-\mathbf{D}^{-1}\mathbf{U}\big)^k$

The matrix $\mathbf{D}$ is an (L-1)-by-(L-1) upper-bidiagonal matrix with upper-diagonal entries equal to $\lambda_1$ and diagonal equal to $-\lambda_1(x,\ldots,x,x_0)$, where $x := (\lambda_1 + \lambda_2(1-z_2)+\ldots+\lambda_M(1-z_M)+1/b_1)/\lambda_1$ and $x_0 := (\lambda_2(1-z_2)+\ldots+\lambda_M(1-z_M)+1/b_1)/\lambda_1$. Thus, it is easy to show that $-\mathbf{D}^{-1}\mathbf{U} = (xb_1\lambda_1)^{-1}\mathbf{L}$, where

$$\mathbf{L} = \begin{pmatrix}
x^{-1} & x^{-2} & x^{-3} & \cdots & x^{-L+3} & x^{-L+3}x_0^{-1} & 0 \\
1 & x^{-1} & x^{-2} & \cdots & x^{-L+4} & x^{-L+4}x_0^{-1} & 0 \\
0 & 1 & x^{-1} & \cdots & x^{-L+5} & x^{-L+5}x_0^{-1} & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & \cdots & 0 & 1 & x_0^{-1} & 0 \\
0 & \cdots & \cdots & \cdots & 0 & x\cdot x_0^{-1} & 0
\end{pmatrix}.$$

For $n\geq 1$, note that the (i,j)-entry of $\mathbf{L}^n$, can be written as $c^n(i,j)x^{-n+i-j}$, $j=1,\ldots,L-3$. We do not consider the (i,j)-entry of $\mathbf{L}^n$ with $j\geq L-2$ since these entries will tend to zero when we will take the limit for $L\to\infty$. The coefficients $c^n(i,j)$ are strictly positive integers for $1\leq i\leq n$ and $1\leq j\leq L-2$, and $n+1\leq i\leq L-1$ and $i-n\leq j\leq L-2$, and zero otherwise. Moreover, the sequence $\{c^n(i,j)\}$ satisfies the following recurrent equation for $n\geq 2$, $1\leq i\leq L-1$ and $1\leq j\leq L-2$,

$$c^n(i,j) = c^n(i,j-1) + c^{n-1}(i,j+1) = \sum_{l=1}^{j+1} c^{n-1}(i,l), \qquad (36)$$

where

$$c^1(i,j) = \begin{cases} 1, & i=1, 1\leq j\leq L-2, \\ 1, & 2\leq i\leq L-1, i-1\leq j\leq L-2, \\ 0, & \text{otherwise.}\end{cases} \qquad (37)$$

The coefficient $c^n(i,j)$ can be interpreted as the number of paths in the directed graph in Figure 1. Especially, $c^n(i,j)$ equals the number of paths from state $i$

in level $l(0)$ to state $j$ in level $l(n)$. Thus by an induction argument, it can be shown that $c^n(i,j)$ has the following solution for $2 \le n < L-2$. That is, for $j = 1, \ldots, L-2$ and $n \ll L$,

$$c^n(1,j) = \left( \begin{array}{c} 2n+j-2 \\ n-1 \end{array} \right) - \left( \begin{array}{c} 2n+j-2 \\ n+j \end{array} \right), \tag{38}$$

for $i = 2, \ldots, n-1$ and $j = 1, \ldots, L-2$,

$$c^n(i,j) = \left( \begin{array}{c} 2n+j-i-1 \\ n-1 \end{array} \right) - \left( \begin{array}{c} 2n+j-i-1 \\ n+j \end{array} \right), \tag{39}$$

for $i = n$ and $j = 1, \ldots, L-2$,

$$c^n(n,j) \;\; = \;\; \left( \begin{array}{c} n+j-1 \\ n-1 \end{array} \right), \tag{40}$$

for $i = 1, \ldots, L-n-1$ and $j = i, \ldots, L-2$,

$$c^n(i+n,j) \;\; = \;\; \left( \begin{array}{c} n+j-i-1 \\ n-1 \end{array} \right), \tag{41}$$

and $c^n(i,j)$ equals zero for $i = n+2, \ldots, L-1$ and $j = 1, \ldots, i-n$.

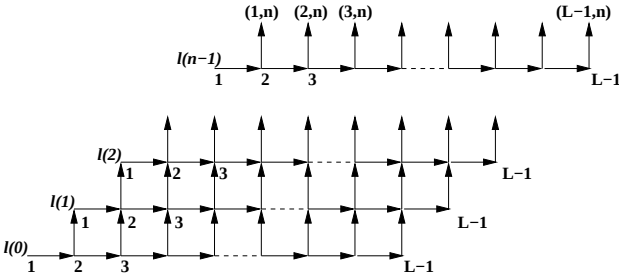


Figure 1: Directed graph for the computation of $c^n(i,j)$.

Finally, we conclude that $e_{i_1}\big(-\mathbf{D}^{-1}\mathbf{U}\big)^k$ is a row vector of size $L-1$ that is equal to $(xb_1\lambda_1)^{-k}\mathbf{L}^k$ with $j$-th element equal to

$$\frac{c^k(i_1,j)}{(\lambda_1 b_1)^k} x^{-2k+i_1-j}, \tag{42}$$

for $j = 1, \ldots, L-3$. Note that since $|x| < 1$, the limit of (42) tends zero for $L \to \infty$.

## 5.2 $\big(\mathbf{D}^{-1}\big)q(z_1)$ and $\big(\mathbf{D}+\mathbf{U}\big)^{-1}e_1$

The matrix $\mathbf{D}$ is a (L-1)-by-(L-1) upper-bidiagonal matrix with upper-diagonal $(\lambda_1, \ldots, \lambda_1)$ and diagonal $-\lambda_1(x, \ldots, x, x_0)$, where $x$ and $x_0$ are defined in Sec-

tion 5.1. Thus,

$$\mathbf{D}^{-1} = \lambda_1^{-1} \begin{pmatrix} x^{-1} & x^{-2} & x^{-2} & \cdots & x^{-L+2} & x^{-L+2}x_0^{-1} \\ 0 & x^{-1} & x^{-2} & \cdots & x^{-L+3} & x^{-L+3}x_0^{-1} \\ 0 & 0 & x^{-1} & \cdots & x^{-L+4} & x^{-L+4}x_0^{-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & x^{-1} & x^{-1}x_0^{-1} \\ 0 & \cdots & \cdots & \cdots & 0 & x_0^{-1} \end{pmatrix}.$$

Using (33), we find that $\mathbf{D}^{-1}q(z_1)$ is an (L-1)-dimensional column vector of $i$-th element, denoted as $d(i)$, equal to

$$\begin{aligned} d(1) &= -\frac{1}{\lambda_1}\Big(z_1x^{-1}\frac{1-(z_1x^{-1})^{L-3}}{x-z_1} + x^{-L+2}x_0^{-1}z_1^{L-2}\Big), \\ d(i) &= -\frac{1}{\lambda_1}\Big(z_1^{i-1}\frac{1-(z_1x^{-1})^{L-1-i}}{x-z_1} + x^{-L+i+1}x_0^{-1}z_1^{L-2}\Big), \end{aligned} \tag{43}$$

for $i = 2, \ldots, L-1$. Note that $|z_1/x| < 1$ which gives that

$$\lim_{L\to\infty} d(1) = -\frac{1}{\lambda_1}\Big(\frac{z_1x^{-1}}{x-z_1}\Big), \tag{44}$$

$$\lim_{L\to\infty} d(i) = -\frac{1}{\lambda_1}\Big(\frac{z_1^{i-1}}{x-z_1}\Big), \tag{45}$$

for all $i < \infty$.

Now we compute $\big(\mathbf{D}+\mathbf{U}\big)^{-1}e_1$. Recall that $\big(\mathbf{D}+\mathbf{U}\big)$ is an (L-1)-by-(L-1) tridiagonal matrix with upper-diagonal entries equal $\lambda_1$, diagonal $-\lambda_1(x, \ldots, x, x_0)$ and lower-diagonal entries $1/b_1$. Therefore, $\big(\mathbf{D}+\mathbf{U}\big)$ is equal to the matrix $\mathbf{Q_T}(\mathbf{z})$ in (22) with $\alpha_1 = 0$. We note that the inverse of $\mathbf{Q_T}(\mathbf{z})$ was computed in (24), thus using these results we find that $\big(\mathbf{D}+\mathbf{U}\big)^{-1}e_1$ is a column vector equal to $\big(p(1), \ldots, p(L-1)\big)^T$ with the $i$-th entry that is given by

$$\begin{aligned} p(i) &:= t(i,1)^*_T - \lambda_1 \frac{t(i,L-1)^*_T t(L-1,1)^*_T}{1+\lambda_1 t(L-1,L-1)^*_T}, \\ &= -b_1\frac{y_{11}^L y_{21}^{i-L} - y_{11}^i}{y_{11}^L y_{21}^{-L} - 1} + \lambda_1 b_1^2 \frac{y_{11}-y_{21}}{y_{21}^{-L} - y_{11}^{-L}} \times \\ &\quad \frac{y_{11}^i - y_{21}^i}{y_{11}^L - y_{11}^{L-1} - y_{21}^L + y_{21}^{L-1}}. \end{aligned} \tag{46}$$

where $y_{11}$ and $y_{21}$ are the distinct roots of

$$\lambda_1 y^2 - s_1^* y + 1/b_1, \tag{47}$$

where $s_1^* := \lambda_1 + 1/b_1 + \sum_{m=2}^M \lambda_m(1-z_m)$. Note that in this case $|y_{11}| \le 1 < |y_{21}|$, so that we may find that,

$$\lim_{L\to\infty} p(i) = -b_1 y_{11}^i, \tag{48}$$

for all $i < \infty$.

### 5.3 Limit of $\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big]$ for $L \to \infty$

Plugging $e_{i_1}\big(\mathbf{D}^{-1}\mathbf{U}\big)^k$, $\big(\mathbf{D}^{-1}\big)q(z_1)$ and $\big(\mathbf{D}+\mathbf{U}\big)^{-1}e_1$ into $\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big]$ and taking the limit for $L \to \infty$ gives that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] \;=\; -1/b_1 z_2^{i_2} \dots z_M^{i_M} S,$$

where,

$$\begin{aligned} S \;:=\; & \frac{b_1 c^{k-1}(i_1,1)}{(\lambda_1 b_1)^k} x^{-2k+i_1} - b_1 y_{11}^{i_1} - \frac{b_1 x^{-2k+i_1+1}}{(\lambda_1 b_1)^k (x-z_1)} \sum_{j=1}^{\infty} c^{k-1}(i_1,j)\Big(\frac{z_1}{x}\Big)^{j-1} \\ & + \frac{b_1 x^{-2k+i_1}}{(\lambda_1 b_1)^k} \sum_{j=1}^{\infty} c^k(i_1,j)\Big(\frac{y_{11}}{x}\Big)^j, \end{aligned} \tag{49}$$

for $k \geq 2$.

Due to the complexity of the analysis for an arbitrary $k$, we will restrict ourselves to the 1-limited and 2-limited disciplines.

**1-limited**. First take the limits of $d(i)$ and $p(i)$ in (45) and (48), then plugging $k = 1$ into (34) and (35) gives that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] \;=\; \frac{z_2^{i_2} \dots z_M^{i_M}}{\lambda_1 b_1 x}\Big(\frac{z_1}{x-z_1} + 1\Big), \tag{50}$$

for $i_1 = 1$, and

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] \;=\; z_2^{i_2} \dots z_M^{i_M}\Big(\frac{z_1^{i_1-1}}{\lambda_1 b_1 (x-z_1)}\Big), \tag{51}$$

for $i_1 = 2, 3, \dots$. Unconditioning on $\mathbf{N}_1^b = (\mathbf{N}_{11}^b, \dots, \mathbf{N}_{M1}^b)$, we find that

$$\begin{aligned} \mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big] = & \frac{1/b_1 z_1^{-1}}{1/b_1 + \lambda_1(1-z_1) + \lambda_2(1-z_2)} \mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^b}\big] + \\ & \Big(1 - \frac{1/b_1 z_1^{-1}}{1/b_1 + \lambda_1(1-z_1) + \lambda_2(1-z_2)}\Big) \mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^b}\big]\Big|_{z_1=0}. \end{aligned} \tag{52}$$

**2-limited**. Plugging $k = 2$ in (49) gives that p.g.f. of $\mathbf{N}_1^e$ then gives

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] \;=\; \frac{z_2^{i_2} \dots z_M^{i_M}}{\lambda_1 b_1 x}\Big(\frac{1}{\lambda_1 b_1 (x-z_1)^2} + 1\Big), \tag{53}$$

for $i_1 = 1$, and

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] = z_2^{i_2} \dots z_M^{i_M}\Big(\frac{z_1^{i-2}}{\lambda_1^2 b_1^2 (x-z_1)^2}\Big), \tag{54}$$

for $i = 2, 3, \ldots$. Unconditioning on $\mathbf{N}_1^b = (\mathbf{N}_{11}^b, \ldots, \mathbf{N}_{M1}^b)$, we find that

$$\begin{aligned}\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big] &= \frac{z_1^{-2}}{\lambda_1^2 b_1^2 (x - z_1)^2}\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^b}\big] + \left(1 - \frac{z_1^{-2}}{\lambda_1^2 b_1^2 (x - z_1)^2}\right)\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^b}\big]\Big|_{z_1=0} \\ &\quad + \left(\frac{z_1^{-1}}{\lambda_1 b_1 x} - \frac{z_1^{-2}}{\lambda_1^2 b_1^2 x(x - z_1)}\right)\mathbb{E}\big[\mathbf{1}_{\{N_{11}^b = 1\}}\mathbf{z}^{\mathbf{N}_1^b}\big]. \qquad (55)\end{aligned}$$

**Remark 1** *The results for 1-limited and 2-limited can also be obtained more directly by explicitly conditioning on the number of jobs at the beginning of a server visit to a queue and keeping track how the queue-length evolves. However, our analysis above shows that our tool can also applied to the k-limited discipline for $k \geq 3$.*

**Remark 2 Exhaustive discipline:**. *The k-limited discipline for $k \to \infty$ is equivalent to the exhaustive discipline. Since $\big(-\mathbf{D}^{-1}\mathbf{U}\big)$ is a sub-stochastic matrix with the sum of its row entries strictly smaller than one, the limit $\big(-\mathbf{D}^{-1}\mathbf{U}\big)^k \to 0$ for $k \to \infty$. Therefore, taking the limit in (34) and (35) for $k \to \infty$ and summing these limits give that*

$$\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] = -1/b_1 z_2^{i_2} \ldots z_M^{i_M} e_{i_1}\big(\mathbf{D} + \mathbf{U}\big)^{-1} e_1. \qquad (56)$$

*The limit of $\mathbb{E}_L\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big]$ for $L \to \infty$ then reads*

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e} \mid \mathbf{N}_1^b\big] = y_{11}^{i_1} z_2^{i_2} \ldots z_M^{i_M}. \qquad (57)$$

*Finally, the unconditioning on $\mathbf{N}_1^b$ gives that*

$$\begin{aligned}\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big] &= \mathbb{E}\big[(\mathbf{z}_1^*)^{\mathbf{N}_1^e}\big], \\ \gamma_1^E(\mathbf{z}) &= \beta_1^E(\mathbf{z}_1^*), \qquad (58)\end{aligned}$$

*where $\mathbf{z}_1^* = (y_{11}, z_2, \ldots, z_M)$. Considering a server visit to $Q_i$, an equivalent relation can be derived for $\gamma_i^E(\mathbf{z})$ and $\beta_1^E(\mathbf{z}_i^*)$ as follows*

$$\gamma_i^E(\mathbf{z}) = \beta_i^E(\mathbf{z}_i^*).$$

*Now including $C^{i-1}$, the switch-over time from $Q_{i-1}$ and $Q_i$, it is easy to find that*

$$\gamma_i^E(\mathbf{z}) = \gamma_{i-1}^E(\mathbf{z}_1^*)C^{i-1}(\mathbf{z}_i^*). \qquad (59)$$

*where $\mathbf{z}_i^* := (z_1, \ldots, z_{i-1}, y_{1i}, z_{i+1}, \ldots, z_M)$ and $y_{1i}$ is the root of*

$$\lambda_i y^2 - s_i^* y + 1/b_i, \qquad (60)$$

*with $|y_{1i}| \leq 1$ and where $s_i^* = \lambda_i + 1/b_i + \sum_{m=1, m\neq i}^{M} \lambda_m(1 - z_m)$. Eq. (59) is equivalent to the well-known relation of exhaustive discipline in (see, e.g., [6, Eq. (24)]).*

# 6 Iterative scheme

In this section, we will explain how to obtain the joint queue-length distribution using an iterative scheme. First, let see how to compute $\gamma_i(\mathbf{z})$ as function $\gamma_{i-1}(\mathbf{z})$, where $\mathbf{z} = (z_1, \ldots, z_M)$.

Note that $\gamma_i(\mathbf{z})$ is a function of $\gamma_{i-1}(\mathbf{z})$ and $\gamma_{i-1}(\mathbf{z}_i^*)$ where $\mathbf{z}_i^* = (z_1, \ldots, z_{i-1}, a, z_{i+1} \ldots, z_M)$ with $|z_i| = 1$, $i = 1, \ldots, M$ and $|a| \leq 1$, which is a function of $z_l$ for all $l = 1, \ldots, M$ and $l \neq i$. Since $\gamma_{i-1}(\mathbf{z})$ is a joint p.g.f., the function $\gamma_{i-1}(\mathbf{z})$ is analytic in $z_i$ for all $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M$. Hence, we can write

$$\gamma_i(\mathbf{z}) = \sum_{n=0}^{\infty} g_{in}(z_1, \ldots, z_{i-1}, z_{i+1} \ldots, z_M) z_i^n, \qquad |a| \leq 1,$$

where $g_{in}(.)$ is again an analytic function. From complex function theory, it is well known that

$$\gamma_i(\mathbf{z}_i^*) = \frac{1}{2\pi \mathbf{i}} \oint_C \frac{\gamma_i(\mathbf{z})}{z_i - a} dz_i, \qquad \text{for } |a| \leq 1,$$

where $C$ is the unit circle and $\mathbf{i}^2 = -1$, and furthermore

$$g_{in}(z_1, \ldots, z_{i-1}, z_{i+1} \ldots, z_M) = \frac{1}{2\pi \mathbf{i}} \oint_C \frac{\gamma_i(\mathbf{z})}{z_i^{n+1}} dz_i,$$

where $n = 0, 1, \ldots$ . These formulas show that we only need to know the joint p.g.f. $\gamma_{i-1}(\mathbf{z})$ for all $\mathbf{z}$ with $|z_i| = 1$, to be able to compute $\gamma_i(\mathbf{z})$.

When there is an incurred switch-over time from queue $i-1$ to $i$ the p.g.f. of the joint queue-length at the end of the n-th server visit to $Q_i$, denoted by $\gamma_i^n(\mathbf{z})$, can be computed as a function of $\gamma_{i-1}^n(\mathbf{z})$. The main step is to iterate over all queues in order to express $\gamma_i^{n+1}(\mathbf{z})$ as a function of $\gamma_i^n(\mathbf{z})$. Assuming that the system is in steady-state these two latter quantities should be equal. Thus, starting with an empty system at the first service visit to $Q_i$ and repeating the latter main step one can compute $\gamma_i^2(\mathbf{z})$, $\gamma_i^3(\mathbf{z})$, and so on. This iteration is stopped when $\gamma_i^n(\mathbf{z})$ converges.

# 7 Tandem model

We know that our tool can be applied also for Jackson-like queueing networks with a single server that can serve only one queue at a time. To show this, we will consider the example of a tandem model of $M$ queues in series. $Q_1$ has Poisson arrivals. The service requirement at $Q_i$ is distributed exponentially with mean $1/b_i$. In the model there is only one server serving the queues according to some schedule. The service discipline is either the autonomous-server or the time-limited. Observe that this tandem model is equivalent to polling system with the property that only $Q_1$ has a Poisson arrivals, the departures from $Q_i$ will join $Q_{i+1}$, $i = 1, \ldots, M-1$, and that departures from $Q_M$ leaves the system.

**Autonomous-server**: according to this discipline the server continues the service of a queue until certain exponentially distributed time of rate $\alpha_1$ will elapse. Consider a server visit to $Q_1$ following the same approach in Section 3 we find that the solution is similar to (11) and the matrix involved has the same structure as $\mathbf{A}_2$. For this reason, we find that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big] = \frac{\alpha_1(z_2 - z_1)}{P_1(z_1)}\left(\frac{r_{11}\mathbb{E}\big[(\mathbf{z}_1^*)^{\mathbf{N}_1^b}\big]}{z_2 - r_{11}} - \frac{z_1\mathbb{E}[\mathbf{z}^{\mathbf{N}_1^b}]}{z_2 - z_1}\right), \tag{61}$$

where $\mathbf{z}_1^* := (r_{11}, z_2, \ldots, z_M)$ and $r_{11}$ is the root of $P_1(r) = \lambda_1 r^2 - (\lambda_1 + 1/b_1 + \alpha_1)r + z_2/b_1$ such that $|r_{11}| < 1$. To relate $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^e}\big]$ to $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^b}\big]$ for a server visit to $Q_i$, $i > 1$, we find that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^e}\big] = \frac{\alpha_i(z_{i+1} - z_i)}{P_i(z_i)}\left(\frac{r_{1i}\mathbb{E}\big[(\mathbf{z}_i^*)^{\mathbf{N}_i^b}\big]}{z_{i+1} - r_{1i}} - \frac{z_i\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^b}]}{z_{i+1} - z_i}\right), \tag{62}$$

where $\mathbf{z}_i^* := (z_1, \ldots, z_{i-1}, r_{1i}, z_{i+1} \ldots, z_M)$ and $r_{1i}$ is the root of $P_i(r) = -(\lambda_1(1 - z_1) + 1/b_i + \alpha_i)r + z_{i+1}/b_i$ such that $|r_{1i}| < 1$.

**Time-limited**: according to this discipline the server continues the service of a queue until certain exponentially distributed time of rate $\alpha_1$ will elapse or the queue becomes empty, whichever occurs first. Consider a server visit to $Q_1$ following the same approach in Section 4 we find that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big] = \Big(1 + \frac{\alpha_1 z_1}{P_1(z_1)}\Big)\mathbb{E}\big[(\mathbf{z}_1^*)^{\mathbf{N}_1^b}\big] - \frac{\alpha_1 z_1}{P_1(z_1)}\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^b}\big], \tag{63}$$

where $\mathbf{z}_1^* := (r_{11}, z_2, \ldots, z_M)$ and $r_{11}$ is the root of $P_1(r) = \lambda_1 r^2 - (\lambda_1 + 1/b_1 + \alpha_1)r + z_2/b_1$ such that $|r_{11}| < 1$. To relate $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^e}\big]$ to $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^b}\big]$ for a server visit to $Q_i$, $i > 1$, we find that

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^e}\big] = \Big(1 + \frac{\alpha_i z_i}{P_i(z_i)}\Big)\mathbb{E}\big[(\mathbf{z}_i^*)^{\mathbf{N}_i^b}\big] - \frac{\alpha_i z_i}{P_i(z_i)}\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^b}\big], \tag{64}$$

where $\mathbf{z}_i^* := (z_1, \ldots, z_{i-1}, r_{1i}, z_{i+1} \ldots, z_M)$ and $r_{1i}$ is the root of $P_i(r) = -(\lambda_1(1 - z_1) + 1/b_i + \alpha_i)r + z_{i+1}/b_i$ such that $|r_{1i}| < 1$.

# 8 Discussion and Conclusion

In this paper, we developed a general framework to analyze polling systems with the autonomous-server, the time-limited, and the k-limited service discipline. The analysis of these disciplines is based on the key idea of relating directly the joint queue-length distribution at the beginning and the end of a server visit. In order to do so, we used the theory of absorbing Markov chain. The analysis presented in this paper is restricted to the case of service requirement with exponential distribution. We emphasize can be extended to more general

distribution such as the phase-type distributions. For instance, Eq. 11 holds in the case of phase-type distribution, however, the matrix $\mathbf{A_2}$ becomes a block matrix which is difficult to invert in closed-form.

In this paper we showed that our tool is not restricted only to the disciplines that do not verify the branching property. For example, we analyzed the exhaustive discipline. Moreover, we claim that with an extra effort one can analyze the gated discipline for which there already exist results in the literature.

# References


[1] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J. van Ommeren. A tandem queueing model for delay analysis in disconnected ad hoc networks. *Proc. of ASMTA*, LNCS 5055:189–205, June 2008.

[2] J. Blanc. An algorithmic solution of polling models with limited service disciplines. *IEEE Transactions on Communications*, 40(7):1152–1155, July 1992.

[3] J. Blanc. The power-series algorithm for polling systems with time limits. *Probability in the Engineering and Informational Sciences*, 12:221–237, 1998.

[4] R. de Haan, R. J. Boucherie, and J. van Ommeren. A polling model with an autonomous server. *Research Memorandum 1845, University of Twente*, 2007.

[5] M. Dow. Explicit inverses of Toeplitz and associated matrices. *ANZIAM J.*, 44(E):E185–E215, Jan. 2003.

[6] M. Eisenberg. Queues with periodic service and changeover times. *Operation Research*, 20(2):440–451, 1972.

[7] T. Estermann. *Complex Numbers and Functions*. Oxford University Press, London, 1962.

[8] D. P. Gaver, P. A. Jacobs, and G. Latouche. Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability*, 16:715–731, 1984.

[9] C. Grinstead and J. Snell. *Introduction to Probability*. American Mathematical Society, 1997.

[10] K. Leung. Cyclic-service systems with probabilistically-limited service. *IEEE Journal on Selected Areas in Communications*, 9(2):185–193, 1991.

[11] K. Leung. Cyclic-service systems with non-preemptive time-limited service. *IEEE Transactions on Communications*, 42(8):2521–2524, 1994.

[12] H. Levy and M. Sidi. Polling systems: Applications, modeling, and optimization. *TOC*, 38(10), Oct. 1990.

[13] M. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins University Press, 1981.

[14] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[15] J. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(10):409–429, 1993.

[16] H. Takagi. Analysis and application of polling models. In *Performance Evaluation: Origins and Directions, LNCS 1769*, pages 423–442, Berlin, Germany, 2000. Springer-Verlag.